第 3 章
概率密度函数的估计

3.1 引言

在上一章最后的讨论中已经提到，贝叶斯决策的基础是概率密度函数的估计，即根据一定的训练样本来估计统计决策中用到的先验概率 $P(\omega_i)$ 和类条件概率密度 $p(\boldsymbol{x}|\omega_i)$。其中，先验概率的估计比较简单，通常只需根据大量样本计算出各类样本在其中所占的比例，或者根据对所研究问题的领域知识事先确定。因此，本章重点介绍类条件概率密度的估计问题。

这种首先通过训练样本估计概率密度函数，再用统计决策进行类别判定的方法称作基于样本的两步贝叶斯决策。

这样得到的分类器性能与第 2 章理论上的贝叶斯分类器有所不同。我们希望当样本数目 $N\to\infty$ 时，基于样本的分类器能收敛于理论上的结果。为做到这一点，实际只要说明 $N\to\infty$ 时，估计的 $\hat{p}(\boldsymbol{x}|\omega_i)$ 和 $\hat{P}(\omega_i)$ 收敛于 $p(\boldsymbol{x}|\omega_i)$ 和 $P(\omega_i)$。这在统计学中可通过对估计量性质的讨论来解决。

在监督学习中，训练样本的类别是已知的，而且假定各类样本只包含本类的信息，这在多数情况下是正确的。因此，我们要做的是利用同一类的样本来估计本类的类条件概率密度。为了讨论方便，在本章后面的论述中，除了特别说明外，一概假定所有样本都是来自同一类，不再标出类别标号。由于概率密度估计是统计推断中的一个重要内容，有很多专门的教科书介绍它，所以在本书中只扼要地介绍其中的基本思想和主要方法，并不深入地展开论述。

概率密度函数的估计方法分为两大类：参数估计(parametric estimation)与非参数估计(nonparametric estimation)。

参数估计中，已知概率密度函数的形式，但其中部分或者全部参数未知，概率密度函数

的估计问题就是用样本来估计这些参数。主要方法又有两类：最大似然估计和贝叶斯估计，两者在很多实际情况下结果接近，但从概念上它们的处理方法是不同的。

非参数估计，就是概率密度函数的形式也未知，或者概率密度函数不符合目前研究的任何分布模型，因此不能仅仅估计几个参数，而是用样本把概率密度函数数值化地估计出来。

参数估计是统计推断的基本问题之一，在讨论具体问题之前先介绍几个参数估计中的基本概念。

(1) 统计量。样本中包含着总体的信息，希望通过样本集把有关信息抽取出来，就是说针对不同要求构造出样本的某种函数，这种函数在统计学中称为统计量。

(2) 参数空间。如上所述，在参数估计中，总是假设总体概率密度函数的形式已知，而未知的仅是分布中的几个参数，将未知参数记为 θ，在统计学中，将总体分布未知参数 θ 的全部可容许值组成的集合称为参数空间，记为 Θ。

(3) 点估计、估计量和估计值。点估计问题就是要构造一个统计量 $d(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$ 作为参数 θ 的估计 $\hat{\theta}$，在统计学中称 $\hat{\theta}$ 为 θ 的估计量。如果 $\boldsymbol{x}_1^{(i)},\cdots,\boldsymbol{x}_N^{(i)}$ 是属于类别 ω_i 的几个样本观察值，代入统计量 d 就得到对于第 i 类的 $\hat{\theta}$ 的具体数值，这个数值在统计学中称为 θ 的估计值。

(4) 区间估计。除点估计外，还有另一类估计，它要求用区间 (d_1,d_2) 作为 θ 可能取值范围的一种估计。这个区间称为置信区间，这类估计问题称为区间估计。

本章要求估计总体分布的具体参数，显然这是点估计问题。我们将介绍两种主要的点估计方法——最大似然估计和贝叶斯估计，它们都能得到相应的估计值。当然评价一个估计的“好坏”，不能按一次抽样结果得到的估计值与参数真值 θ 的偏差大小来确定，而必须从平均的和方差的角度出发进行分析。为了表示这种偏差，统计学中做了很多关于估计量性质的定义。

在数理统计中，用来判断估计好坏的常用标准是无偏性、有效性和一致性，这是本章介绍的估计方法的基本出发点。如果参数 θ 的估计量 $\hat{\theta}(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$ 的数学期望等于 θ，则称估计是无偏的；如果当样本数趋于无穷时估计才具有无偏性，则称为渐近无偏。如果一种估计的方差比另一种估计的方差小，则称方差小的估计更有效。而如果对于任意给定的正数 ε，总有

$$\lim_{n\to\infty} P(|\hat{\theta}_n-\theta|>\varepsilon)=0$$

则称 $\hat{\theta}$ 是 θ 的一致估计。显然，无偏性、有效性都只是说明对于多次估计来说，估计量能以较小的方差平均地表示其真实值，并不能保证具体的一次估计的性能；而一致性则保证当样本数无穷多时，每一次的估计量都将在概率意义上任意地接近其真实值。

在统计学中常见的一些典型分布形式并不总是能够拟合所有实际数据的分布，此外，在很多实际问题中还会遇到多峰分布的情况，在另外一些情况下我们可能无法事先判断数据的分布情况。在这些情况下，都需要直接利用样本去非参数地估计概率密度。本章将介绍其中最基本的直方图法、k_N 近邻法和 Parzen 窗法。

3.2 最大似然估计

3.2.1 最大似然估计的基本原理

在最大似然估计(maximum likelihood estimation)中,我们做以下基本假设:

(1) 我们把要估计的参数记作 θ,它是确定但未知的量(多个参数时是向量),这与把它看作随机量的方法是不同的。

(2) 每类的样本集记作 $\mathcal{X}_i$,$i=1,2,\cdots,c$,其中的样本都是从密度为 $p(\boldsymbol{x}|\omega_i)$的总体中独立抽取出来的,即所谓满足独立同分布条件。

(3) 类条件概率密度 $p(\boldsymbol{x}|\omega_i)$具有某种确定的函数形式,只是其中的参数 θ 未知。比如在 x 是一维正态分布 $N(\mu,\sigma^2)$时,未知的参数就可能是$\boldsymbol{\theta}=[\mu,\sigma^2]^{\mathrm{T}}$,对不同类别的参数可以记作 θ_i,为了强调概率密度中待估计的参数,也可以把 $p(\boldsymbol{x}|\omega_i)$写作 $p(\boldsymbol{x}|\omega_i,\theta_i)$或 $p(\boldsymbol{x}|\theta_i)$。

(4) 各类样本只包含本类的分布信息,也就是说,不同类别的参数是独立的,这样就可以分别对每一类单独处理。

在这些假设的前提下,我们就可以分别处理 c 个独立的问题。即,在一类中独立地按照概率密度 $p(\boldsymbol{x}|\theta)$抽取样本集 $\mathcal{X}$,用 $\mathcal{X}$ 来估计出未知参数 θ。

设样本集包含 N 个样本,即

$$\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\} \tag{3-1}$$

由于样本是独立地从 $p(\boldsymbol{x}|\theta)$中抽取的,所以在概率密度为 $p(\boldsymbol{x}|\theta)$时获得样本集 $\mathcal{X}$ 的概率即出现 $\mathcal{X}$ 中的各个样本的联合概率是

$$l(\theta)=p(\mathcal{X}|\theta)=p(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N|\theta)=\prod_{i=1}^{N}p(\boldsymbol{x}_i|\theta) \tag{3-2}$$

这个概率反映了在概率密度函数的参数是 θ 时,得到式(3-1)中这组样本的概率。现在因为已经得到了式(3-1)的样本集,而 θ 是不知道的,式(3-2)就成为 θ 的函数,它反映的是在不同参数取值下取得当前样本集的可能性,因此称作参数 θ 相对于样本集 $\mathcal{X}$ 的似然函数(likelihood function)。式(3-2)右边乘积中的每一项 $p(\boldsymbol{x}_i|\theta)$就是 θ 相对于每一个样本的似然函数。

似然函数 $l(\theta)$给出了从总体中抽出 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$ 这样 N 个样本的概率。为了便于解释,暂且假定 θ 是已知的,用 θ_0 表示这个已知值。最可能出现的 N 个样本是使得 $l(\theta)$值为最大的样本 $\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_N$。例如,为了简便,假定 $N=1$,$\boldsymbol{x}$ 为一维且具有以均值为 6,方差为 1 的正态分布,那么“最可能出现的”样本就是 $x'_1=6$,这时似然函数 $l(\theta)=p(x'_1|6,1)=\max_{x\in E^d}p(\boldsymbol{x}|6,1)$。再回过来看,假定 θ 为未知,而我们从一次抽样中得到 N 个样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$,想知道这组样本“最可能”来自哪个密度函数。换句话说,所抽取出的这组样本,来自哪个密度函数(θ 取什么值)的可能性最大?即我们要在参数空间 Θ 中找到一个 θ 值(用 $\hat{\theta}$表示),它能使似然函数 $l(\theta)$极大化。一般来说,使似然函数的值最大的 $\hat{\theta}$是样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$ 的函数,记为 $\hat{\theta}=d(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$,就把 $\hat{\theta}=d(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$叫做 θ 的最大似然估计量。

图 3-1 示意了最大似然估计的基本原理。现在可以给出最大似然估计量的定义。

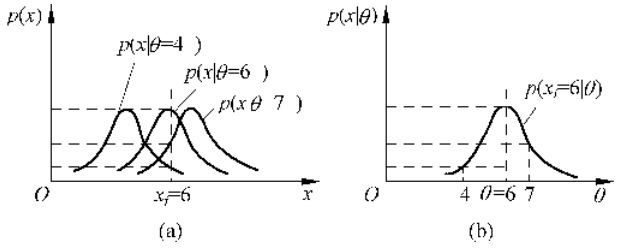

图 3-1 最大似然函数示意图

最大似然估计量：令 $l(\theta)$为样本集 $\mathcal{X}$的似然函数，$\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，如果$\hat{\theta}=d(\mathcal{X})=d(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$是参数空间 Θ 中能使似然函数 $l(\theta)$极大化的 θ 值，那么 $\hat{\theta}$就是 θ 的最大似然估计量，或者记作

$$\hat{\theta} = \arg\max l(\theta) \tag{3-3}$$

其中，arg max 是一种常用的表示方法，表示使后面的函数取得最大值的变量的取值。有时，为了便于分析，还可以定义对数似然函数

$$H(\theta) = \ln l(\theta) = \ln\prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \sum_{i=1}^{N}\ln p(\boldsymbol{x}_i|\theta) \tag{3-4}$$

容易证明，使对数似然函数最大的 θ 值也使似然函数最大。

3.2.2 最大似然估计的求解

现在再来看式(3-2)似然函数公式，虽然公式复杂，但其中的函数形式 $p(\cdot)$是已知的，其中的 $\boldsymbol{x}_i$ 也都是已知的，未知量只有 θ。在似然函数满足连续、可微的条件下，如果 θ 是一维变量，即只有一个待估计参数，其最大似然估计量就是如下微分方程的解

$$\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta} = 0 \tag{3-5}$$

或

$$\frac{\mathrm{d}H(\theta)}{\mathrm{d}\theta} = 0 \tag{3-6}$$

更一般地，当$\boldsymbol{\theta}=[\theta_1,\cdots,\theta_s]^{\mathrm{T}}$ 是由多个未知参数组成的向量时，求解似然函数的最大值就需要对$\boldsymbol{\theta}$的每一维分别求偏导，即用下面的梯度算子

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1},\cdots,\frac{\partial}{\partial\theta_s}\right]^{\mathrm{T}} \tag{3-7}$$

来对似然函数或者对数似然函数求梯度并令其等于零

$$\nabla_\theta l(\theta) = 0 \tag{3-8}$$

或

$$\nabla_\theta H(\boldsymbol{\theta}) = \sum_{i=1}^{N}\nabla_\theta \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) = 0 \tag{3-9}$$

得到 s 个方程，方程组的解就是对数似然函数的极值点。需要注意，在某些情况下，似然函

数可能有多个极值，此时上述方程组可能有多个解，其中使得似然函数最大的那个解才是最大似然估计量。

并不是所有的概率密度形式都可以用上面的方法求得最大似然估计。比如，已知一维随机变量 x 服从均匀分布

$$p(x|\theta)=\begin{cases}\dfrac{1}{\theta_2-\theta_1} & \theta_1<x<\theta_2\\ 0 & \text{其他}\end{cases} \tag{3-10}$$

其中分布的参数 θ_1、θ_2 未知。从总体分布中独立抽取了 N 个样本 $x_1,x_2,\cdots,x_N$，则似然函数为

$$l(\theta)=p(\mathcal{X}|\theta)=\begin{cases}p(x_1,x_2,\cdots,x_N|\theta_1,\theta_2)=\dfrac{1}{(\theta_2-\theta_1)^N}\\ 0\end{cases} \tag{3-11}$$

对数似然函数为

$$H(\theta)=-N\ln(\theta_2-\theta_1) \tag{3-12}$$

通过式(3-9)求

$$\frac{\partial H}{\partial\theta_1}=N\cdot\frac{1}{\theta_2-\theta_1} \tag{3-13}$$

$$\frac{\partial H}{\partial\theta_2}=-N\cdot\frac{1}{\theta_2-\theta_1} \tag{3-14}$$

从式(3-13)、式(3-14)方程组中解出的参数 θ_1 和 θ_2 至少有一个为无穷大，这是无意义的结果。造成这种困难的原因是似然函数在最大值的地方没有零斜率，所以必须用其他方法来找最大值。从式(3-10)看出，当 $\theta_2-\theta_1$ 越小时，则似然函数越大。而在给定一个有 N 个观察值 $x_1,x_2,\cdots,x_N$ 的样本集中，如果用 x' 表示观察值中最小的一个，用 x'' 表示观察值中最大的一个，显然 θ_1 不能大于 x'，θ_2 不能小于 x''，因此 $\theta_2-\theta_1$ 的最小可能值是 $x''-x'$，这时 θ 的最大似然估计量显然是

$$\hat{\theta}_1=x' \tag{3-15}$$

$$\hat{\theta}_2=x'' \tag{3-16}$$

3.2.3 正态分布下的最大似然估计

这里仅以单变量正态分布情况下估计其均值和方差为例来说明最大似然估计的用法。我们知道，单变量正态分布的形式为

$$p(x|\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3-17}$$

其中均值 μ 和方差 σ^2 为未知参数，即我们要估计的参数为 $\boldsymbol{\theta}=[\theta_1,\theta_2]^{\mathrm{T}}=[\mu,\sigma^2]^{\mathrm{T}}$，用于估计的样本仍然是 $\mathcal{X}=\{x_1,x_2,\cdots,x_N\}$。根据式(3-11)，最大似然估计应该是下面方程组的解

$$\nabla_\theta H(\boldsymbol{\theta})=\sum_{k=1}^{N}\nabla_\theta\ln p(x_k|\boldsymbol{\theta})=0 \tag{3-18}$$

从正态分布式(3-17)可以得到

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k-\theta_1)^2 \tag{3-19}$$

分别对两个未知参数求偏导，得到

$$\nabla_\theta \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k-\theta_1) \\ -\frac{1}{2\theta_2}+\frac{1}{2\theta_2^2}(x_k-\theta_1)^2 \end{bmatrix} \tag{3-20}$$

因此，最大似然估计应该是以下方程组的解

$$\begin{cases} \sum_{k=1}^{N}\frac{1}{\hat{\theta}_2}(x_k-\hat{\theta}_1)=0 \\ -\sum_{k=1}^{N}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{N}\frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2}=0 \end{cases} \tag{3-21}$$

容易解得

$$\hat{\mu}=\hat{\theta}_1=\frac{1}{N}\sum_{k=1}^{N}x_k \tag{3-22}$$

$$\hat{\delta}^2=\hat{\theta}_2=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})^2 \tag{3-23}$$

这正是人们经常使用的对均值和方差的估计，它们是对正态分布样本的均值和方差的最大似然估计。

对于多元正态分布，分析原理和上面相同，只是公式略微复杂一些，结论也和单变量情况很相似，即多元正态分布的均值和方差的最大似然估计是

$$\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{3-24}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{3-25}$$

从以上结果可以得出结论：均值向量$\boldsymbol{\mu}$的最大似然估计是样本均值。协方差矩阵$\boldsymbol{\Sigma}$的最大似然估计是 N 个矩阵$(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})^{\mathrm{T}}$的算术平均。由于真正的协方差矩阵是随机矩阵$(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}$的期望值，所以这个结果是非常令人满意的。最大似然估计量是平方误差一致和简单一致估计量，但不一定都是无偏估计量。上例中$\hat{\boldsymbol{\mu}}$是无偏的，而$\hat{\boldsymbol{\Sigma}}$就不是无偏的，$\hat{\boldsymbol{\Sigma}}$的无偏估计为$\frac{1}{N-1}\sum_{k=1}^{N}(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})^{\mathrm{T}}$，其证明作为习题留给读者。

3.3 贝叶斯估计与贝叶斯学习

贝叶斯估计(Bayesian Estimation)是概率密度估计中的另一类主要的参数估计方法，其结果在很多情况下与最大似然法相同或几乎相同，但是两种方法对问题的处理视角是不同的，在应用上也各自有各自的特点。一个根本的区别是，最大似然估计是把待估计的参数当作未知但固定的量，要做的是根据观测数据估计这个量的取值；而贝叶斯估计则把待估

计的参数本身也看作是随机变量，要做的是根据观测数据对参数的分布进行估计，除了观测数据外，还可以考虑参数的先验分布。贝叶斯学习(Bayesian Learning)则是把贝叶斯估计的原理用于直接从数据对概率密度函数进行迭代估计。

3.3.1 贝叶斯估计

可以把概率密度函数的参数估计问题看作一个贝叶斯决策问题，但是这里要决策的不是离散的类别，而是参数的取值，是在连续的空间里做决策。

把待估计参数 θ 看作具有先验分布密度 $p(\theta)$ 的随机变量，其取值与样本集 $\mathcal{X}$ 有关，我们要做的是根据样本集 $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 估计最优的 θ(记作 θ^*)。

在用于分类的贝叶斯决策中，最优的条件可以是最小错误率或者最小风险。在这里，对连续变量 θ，我们假定把它估计为 $\hat{\theta}$ 所带来的损失为 $\lambda(\hat{\theta},\theta)$，也称作损失函数。

设样本的取值空间是 E^d，参数的取值空间是 Θ，那么，当用 $\hat{\theta}$ 来作为估计时总期望风险就是

$$\begin{aligned} R&=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\boldsymbol{x},\theta)\mathrm{d}\theta\mathrm{d}\boldsymbol{x}\\ &=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\theta\mathrm{d}\boldsymbol{x} \end{aligned} \tag{3-26}$$

我们定义在样本 $\boldsymbol{x}$ 下的条件风险为

$$R(\hat{\theta}|\boldsymbol{x})=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta|\boldsymbol{x})\mathrm{d}\theta \tag{3-27}$$

那么，式(3-26)就可以写成

$$R=\int_{E^d}R(\hat{\theta}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{3-28}$$

现在的目标是对期望风险求最小。与贝叶斯分类决策时相似，这里的期望风险也是在所有可能的 $\boldsymbol{x}$ 情况下的条件风险的积分，而条件风险又都是非负的，所以求期望风险最小就等价于对所有可能的 $\boldsymbol{x}$ 求条件风险最小。在有限样本集合 $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 的情况下，我们所能做的就是对所有的样本求条件风险最小，即

$$\theta^*=\arg\min_{\hat{\theta}}R(\hat{\theta}|\mathcal{X})=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta|\mathcal{X})\mathrm{d}\theta \tag{3-29}$$

在决策分类时，需要事先定义决策表即损失表，而连续情况下，需要定义损失函数。最常用的损失函数是平方误差损失函数，即

$$\lambda(\hat{\theta},\theta)=(\theta-\hat{\theta})^2 \tag{3-30}$$

可以证明，如果采用平方误差损失函数，则在样本 $\boldsymbol{x}$ 条件下 θ 的贝叶斯估计量 θ^* 是在给定 $\boldsymbol{x}$ 下 θ 的条件期望，即

$$\theta^*=E[\theta|\boldsymbol{x}]=\int_{\Theta}\theta p(\theta|\boldsymbol{x})\mathrm{d}\theta \tag{3-31}$$

同样，在给定样本集 $\mathcal{X}$ 下，θ 的贝叶斯估计量是

$$\theta^*=E[\theta|\mathcal{X}]=\int_{\Theta}\theta p(\theta|\mathcal{X})\mathrm{d}\theta \tag{3-32}$$

这样，在最小平方误差损失函数下，贝叶斯估计的步骤是：

(1) 根据对问题的认识或者猜测确定 θ 的先验分布密度 $p(\theta)$。

(2) 由于样本是独立同分布的，而且已知样本密度函数的形式 $p(\boldsymbol{x}|\theta)$，可以形式上求出样本集的联合分布为

$$p(\mathscr{X}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) \tag{3-33}$$

其中 θ 是变量。

(3) 利用贝叶斯公式求 θ 的后验概率分布

$$p(\theta|\mathscr{X}) = \frac{p(\mathscr{X}|\theta)p(\theta)}{\int_{\Theta} p(\mathscr{X}|\theta)p(\theta)\mathrm{d}\theta} \tag{3-34}$$

(4) 根据式(3-32)，θ 的贝叶斯估计量是

$$\theta^* = \int_{\Theta} \theta p(\theta|\mathscr{X})\mathrm{d}\theta \tag{3-35}$$

在贝叶斯估计中，样本的概率密度函数 $p(\boldsymbol{x}|\theta)$ 的形式是已知的，参数的先验分布密度 $p(\theta)$ 只有在某些特殊形式下才能使式(3-34)的后验概率形式上方便计算。特别地，对于给定的概率密度函数 $p(\boldsymbol{x}|\theta)$ 模型，如果先验密度 $p(\theta)$ 能够使参数的后验分布 $p(\theta|\mathscr{X})$ 具有与 $p(\boldsymbol{x}|\theta)$ 相同的形式，则这样的先验密度函数形式称作与概率模型 $p(\boldsymbol{x}|\theta)$ 共轭(conjugate)。在 Bernardo 和 Smith 的教材(Bernardo J M and Smith A F M. *Bayesian Theory*. Chichester: Wiley, 1994)中给出了一些常用的共轭先验概率密度模型的例子，实际中最常用的是在 $p(\boldsymbol{x}|\theta)$ 为正态分布时 $p(\theta)$ 也为正态分布。

应注意到，我们本来的目的并不是估计概率密度参数，而是估计样本的概率密度函数 $p(\boldsymbol{x}|\mathscr{X})$ 本身，因为假定概率密度函数的形式已知，才转化为估计密度函数中的参数的问题。实际上，在上面介绍的贝叶斯估计框架下，从式(3-34)得到了参数的后验概率后就可以不必求对参数的估计，而是直接得到样本的概率密度函数

$$p(\boldsymbol{x}|\mathscr{X}) = \int_{\Theta} p(\boldsymbol{x}|\theta)p(\theta|\mathscr{X})\mathrm{d}\theta \tag{3-36}$$

可以这样直观地理解式(3-36)：参数 θ 是随机变量，它有一定的分布，而要估计的概率密度 $p(\boldsymbol{x}|\mathscr{X})$ 就是所有可能的参数取值下的样本概率密度的加权平均，而这个加权就是在观测样本下估计出的参数 θ 的后验概率。在式(3-34)给出的参数分布估计中，决定分布形状的是 $p(\mathscr{X}|\theta)p(\theta)$，即

$$p(\theta|\mathscr{X}) \sim p(\mathscr{X}|\theta)p(\theta) \tag{3-37}$$

分母只是对估计出的分布的归一化因子，保证概率密度函数下的积分为 1。可以看到，$p(\theta|\mathscr{X})$ 是由两项决定的，一项就是上一节定义的似然函数 $p(\mathscr{X}|\theta)$，它反映了在不同参数取值下得到观测样本的可能性；另一项是参数取值的先验概率 $p(\theta)$，它反映了对参数分布的先验知识或者主观猜测。

极端情况下，如果完全没有先验知识，即认为 $p(\theta)$ 是均匀分布，则 $p(\theta|\mathscr{X})$ 就完全取决于 $p(\mathscr{X}|\theta)$。与最大似然估计不同的是，这里并没有直接把似然函数最大或者是后验概率最大的值拿来当作对样本概率密度参数的估计，而是根据把所有可能的参数值都考虑进来，用它们的似然函数作为加权来平均出一个对参数的估计(即式(3-35))或者对样本概率密度

函数的估计(式(3-36))。

另一个极端情况是,如果先验知识非常强,$p(\theta)$就是在某一特定取值 θ_0 上的一个脉冲函数(即 $p(\theta_0)=\delta(\theta_0)$),则由式(3-37)可知,除非在 θ_0 的似然函数为 0,否则最后的估计就是 θ_0,样本不再起作用。

通常情况下,$p(\theta)$不是均匀分布也不是脉冲函数,则参数的后验概率就受似然函数和先验概率的共同作用。一种常见的情况是,似然函数在其最大值 $\theta=\hat{\theta}$ 附近会有一个尖峰,那么如果先验概率在最大似然估计处不为零且变化比较平缓,则参数的后验概率 $p(\theta|\mathcal{X})$就会集中在 $\theta=\hat{\theta}$ 附近,此时式(3-35)得到的贝叶斯估计就与最大似然估计接近,式(3-36)估计出的样本密度基本上也就是在最大似然估计下的样本密度。

3.3.2 贝叶斯学习

现在来考虑更为一般的情况,即根据观测样本用式(3-35)来估计样本概率密度函数的参数。为了反映样本的数目,把样本集重新记作 $\mathcal{X}^N=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,式(3-35)重写如下

$$\theta^* = \int_{\Theta}\theta p(\theta|\mathcal{X}^N)\mathrm{d}\theta \tag{3-38}$$

其中

$$p(\theta|\mathcal{X}^N) = \frac{p(\mathcal{X}^N|\theta)p(\theta)}{\int_{\Theta}p(\mathcal{X}^N|\theta)p(\theta)\mathrm{d}\theta} \tag{3-39}$$

当 $N>1$ 时,有

$$p(\mathcal{X}^N|\theta) = p(\boldsymbol{x}_N|\theta)p(\mathcal{X}^{N-1}|\theta) \tag{3-40}$$

把它代入式(3-39),可以得到如下的递推公式

$$p(\theta|\mathcal{X}^N) = \frac{p(\boldsymbol{x}_N|\theta)p(\theta|\mathcal{X}^{N-1})}{\int_{p}(\boldsymbol{x}_N|\theta)p(\theta|\mathcal{X}^{N-1})\mathrm{d}\theta} \tag{3-41}$$

为了形式统一,把先验概率记作 $p(\theta|\mathcal{X}^0)=p(\theta)$,表示在没有样本情况下的概率密度估计。根据式(3-41),随着样本数的增加,可以得到一系列对概率密度函数参数的估计

$$p(\theta),\quad p(\theta|\boldsymbol{x}_1),\quad p(\theta|\boldsymbol{x}_1,\boldsymbol{x}_2),\cdots,\quad p(\theta|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N),\cdots \tag{3-42}$$

称作递推的贝叶斯估计。如果随着样本数的增加,式(3-42)的后验概率序列逐渐尖锐,逐步趋向于以 θ 的真实值为中心的一个尖峰,当样本无穷多时收敛于在参数真实值上的脉冲函数,则这一过程称作贝叶斯学习。

此时,用式(3-36)估计的样本概率密度函数也逼近真实的密度函数

$$p(\boldsymbol{x}|\mathcal{X}^{N\to\infty}) = p(\boldsymbol{x}) \tag{3-43}$$

3.3.3 正态分布时的贝叶斯估计

下面以最简单的一维正态分布模型为例来说明贝叶斯估计的应用。假设模型的均值 μ 是待估计的参数,方差为 σ^2 为已知,我们可以把分布密度写为

$$p(x|\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{3-44}$$

假定均值 μ 的先验分布也是正态分布，其均值为 μ_0、方差为 σ_0^2，即

$$p(\mu)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \tag{3-45}$$

用式(3-34)来对均值 μ 进行估计

$$p(\mu|\mathcal{X})=\frac{p(\mathcal{X}|\mu)p(\mu)}{\int_{\Theta}p(\mathcal{X}|\mu)p(\mu)\mathrm{d}\mu} \tag{3-46}$$

已经知道，这里的分母只是用来对估计出的后验概率进行归一化的常数项，可以暂时不考虑。现在来计算式(3-46)右边的分子部分。

$$\begin{aligned}p(\mathcal{X}|\mu)p(\mu)&=p(\mu)\prod_{i=1}^{N}p(x_i|\mu)\\&=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\right)\end{aligned}$$

把所有不依赖于 μ 的量都写入一个常数中，上式可以整理为

$$p(\mathcal{X}|\mu)p(\mu)=\alpha\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right) \tag{3-47}$$

可见 $p(\mu|\mathcal{X})$也是一个正态分布，可以得到

$$p(\mu|\mathcal{X})=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right) \tag{3-48}$$

其中的参数满足

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2} \tag{3-49}$$

$$\mu_N=\sigma_N^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right) \tag{3-50}$$

进一步整理后得

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 \tag{3-51}$$

$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2} \tag{3-52}$$

其中，$m_N=\frac{1}{N}\sum_{i=1}^{N}x_i$ 是所有观测样本的算术平均。

所以，贝叶斯估计告诉我们，待估计的样本密度函数的均值参数服从均值为 μ_N、方差为 σ_N^2 的正态分布。显然，可以用式(3-35)得到参数的贝叶斯估计值，即

$$\hat{\mu}=\int\mu p(\mu|\mathcal{X})\mathrm{d}\mu=\int\frac{\mu}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right)\mathrm{d}\mu=\mu_N \tag{3-53}$$

在式(3-51)中，正态分布下贝叶斯估计的结果是由两项组成的，一项是样本的算术平均，另一项是对均值的先验认识。当样本数目趋于无穷大时，第一项的系数趋于 1 而第二项的系数趋于 0，即估计的均值就是样本的算术平均。这与最大似然估计是一致的。当样本

数目有限时，如果先验知识非常确定，那么先验分布的方差 σ_0^2 就很小，此时第一项的系数就很小，而第二项的系数接近 1，估计主要由先验知识来决定。一般情况下，均值的贝叶斯估计是在样本算术平均与先验分布均值之间进行加权平均。

从这里可以看到，贝叶斯估计的优势不但在于使用样本中提供的信息进行估计，而且能够很好地把关于待估计参数的先验知识融合进来，并且能够根据数据量大小和先验知识的确定程度来调和两部分信息的相对贡献，这在很多实际问题中是非常有用的。

在得到式(3-48)的后验分布后，我们也可以直接用式(3-36)求出样本的密度函数

$$p(\boldsymbol{x}|\mathcal{X}) = \frac{1}{\sqrt{2\pi}\ \sqrt{\sigma^2+\sigma_N^2}}\exp\left\{-\frac{1}{2}\left(\frac{\boldsymbol{x}-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right\} \sim N(\mu_N, \sigma^2+\sigma_N^2) \tag{3-54}$$

其中，μ_N、σ_N^2 仍然如式(3-52)、式(3-53)。可以看到，虽然我们的条件是已知方差 σ^2，但是由于均值是估计值 μ_N，贝叶斯估计得到的分布密度函数方差增加了，变成 $\sigma^2+\sigma_N^2$，而所增加的项 σ_N^2 在当样本趋于无穷大时趋于零。

3.3.4 其他分布的情况

需要说明的是，在一般情况下，在求出参数的后验概率分布 $p(\theta|\mathcal{X})$ 后，计算式(3-35)的数学期望和式(3-36)的积分并不是非常容易的，对于某些概率模型甚至会非常困难。在这种情况下，比较简单的做法是直接根据 $p(\theta|\mathcal{X})$ 选取一个参数值作为估计，比如选择后验概率最大的参数值，但在很多分布下最大值与数学期望的差距可能会很大。一种更完善的做法是，用所谓吉布斯采样(Gibbs Sampling)等方法来对参数的后验概率分布 $p(\theta|\mathcal{X})$ 进行随机采样，用采样得到的参数的算术平均来估算式(3-35)的数学期望，而这种采样并不需要计算式(3-34)中的分母部分(在很多分布下这一归一化因子的计算并不容易)，而是可以根据与 $p(\theta|\mathcal{X})$ 成正比的 $p(\mathcal{X}|\theta)p(\theta)$ 进行。有关这方面的内容请参阅 Andrew Webb 的教材 *Statistical Pattern Recognition* 中有关内容或者关于马尔可夫链蒙特卡罗 MCMC (Markov Chain Monte Carlo)的有关教材或专著。

3.4 概率密度估计的非参数方法

最大似然方法和贝叶斯方法都属于参数化的估计方法，要求待估计的概率密度函数形式已知，只是利用样本来估计函数中的某些参数。但是，在很多情况下，我们对样本的分布并没有充分的了解，无法事先给出密度函数的形式，而且有些样本分布的情况也很难用简单的函数来描述。在这种情况下，就需要非参数估计，即不对概率密度函数的形式作任何假设，而是直接用样本估计出整个函数。当然，这种估计只能是用数值方法取得，无法得到完美的封闭函数形式。从另外的角度来看，概率密度函数的参数估计实际是在指定的一类函数中选择一个函数作为对未知函数的估计，而非参数估计则可以看作是从所有可能的函数中进行的一种选择。